Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 11-025

A regression model for predicting optimal purchase timing for airline tickets

William Groves and Maria Gini

October 18, 2011

A regression model for predicting optimal purchase timing for airline tickets

William Groves and Maria Gini
Department of Computer Science and Engineering, University of Minnesota
{groves, gini}@cs.umn.edu

Abstract

Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. This paper presents a model for computing expected future prices and reasoning about the risk of price changes. The proposed model is used to predict the future expected minimum price of all available flights on specific routes and dates based on a corpus of historical price quotes. Also, we apply our model to predict prices of flights with specific desirable properties such as flights from a specific airline, non-stop only flights, or multi-segment flights. By comparing models with different target properties, buyers can determine the likely cost of their preferences. We present the expected costs of various preferences for two high-volume routes. Performance of the prediction models presented is achieved by including instances of time-delayed features, by imposing a class hierarchy among the raw features based on feature similarity, and by pruning the classes of features used in prediction based on in-situ performance. Our results show that purchase policy guidance using these models can lower the average cost of purchases in the 2 month period prior to a desired departure. The proposed method compares favorably with a deployed commercial web site providing similar purchase policy recommendations.

1 Introduction

Adversarial risk in the airline ticket domain exists in two contexts: the adversarial relationship between buyers and sellers, and the competitive relationships that exist between multiple airlines providing the equivalent service. Buyers are often seeking the lowest price on their travel, while sellers are seeking to keep overall revenue as high as possible to maximize profit. Simultaneously, each seller must consider the price movements of its competitors to ensure that its prices remain sufficiently competitive to achieve sufficient (but not too high) demand. It is impossible to effectively address the problem of optimizing decision making from the buyer's point of view without also considering both types of adversarial relationships.

Sellers (airlines) make significant long term investments in fixed infrastructure (airports, repair facilities), hardware (planes), and route contracts. The specific details of these long term decisions are intended to roughly match expected demand but often do not match exactly. Dynamic setting of prices is the mechanism that airlines use to increase the matching between their individual supply and demand profile in order to attain the greatest revenue.

A central challenge in the airline ticket purchasing domain is the information asymmetry that exists between buyers and sellers. Airlines have the ability to mine significant databases of historical sales data to develop models for expected future demand for each flight. Demand for a specific flight is likely to vary over time and will also vary based on the pricing strategy adopted by the airline. For buyers, it is generally best to buy far in advance of a flight's departure because the prices tend to increase dramatically as the departure date approaches. But, airlines often violate this principle and adjust prices downward to increase sales.

We make two novel contributions in this work: (1) a method of automated optimal feature set generation from the data that leverages a hierarchicalization of the available features to efficiently compute a feature set is proposed; (2) the addition of time-delayed observations to the feature vector fed to the machine learning algorithm is performed. This allows anticipation of trends and more complex relationships between variables. For instance, we address pricing behaviors up to and beyond 60 days prior to departure, and we consider purchasing a flight on any airline for a specific date and city pair (previous work only considers the cost of a specific pair of flight numbers from two specific airlines).

These ideas are then experimentally applied to prediction in the real-world airline ticket purchase domain. This paper presents models that also accommodate preferences of passengers about the number of stops in the itinerary or the specific airline to use. We believe this prediction task is both a more difficult task and generates models that are more useful for actual airline passengers.

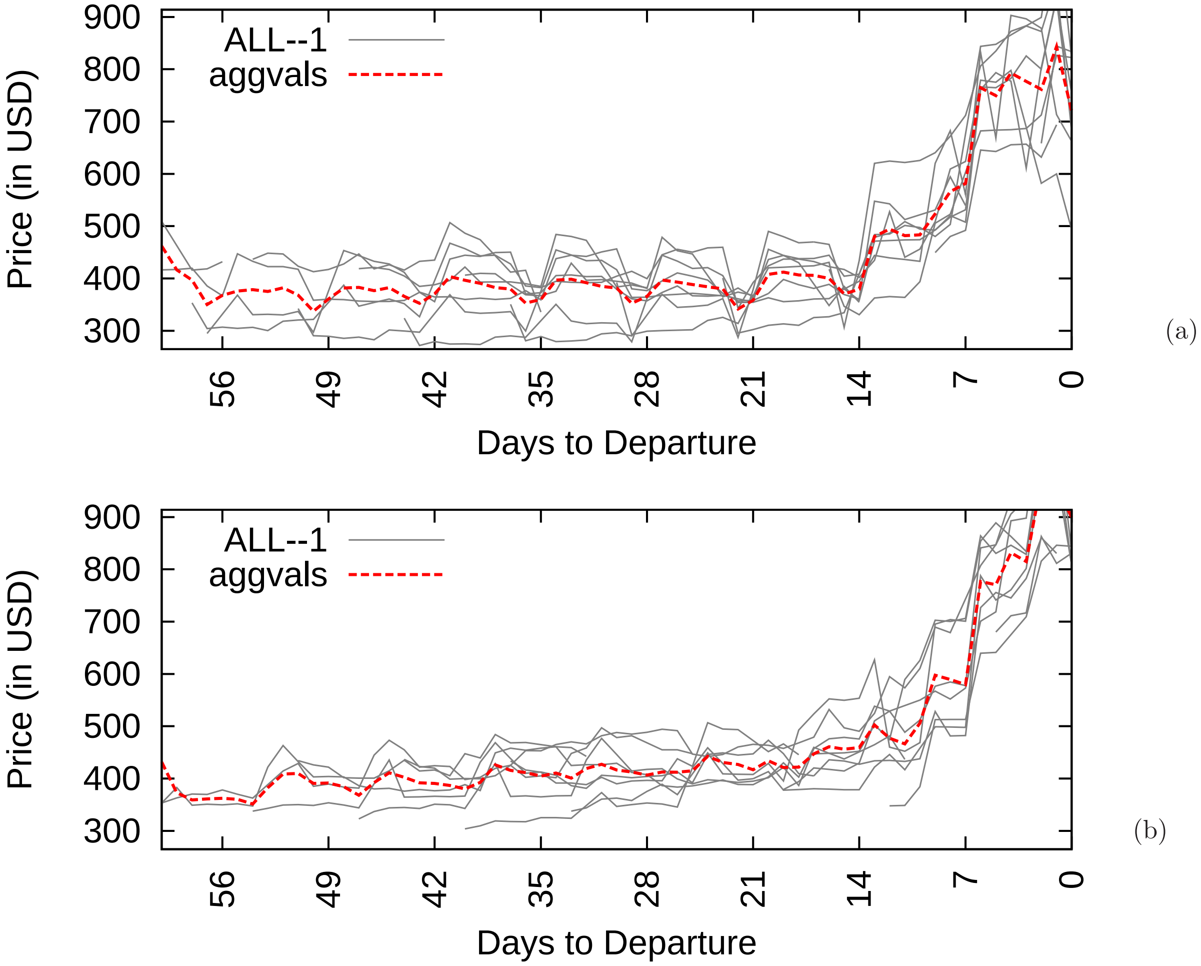

Figure 1: Mean lowest price offered by all airlines for MSP to NYC 5-day round trip flights having (a) Thursday departure and Tuesday return, or (b) Monday departure and Friday return itineraries. Each solid line series indicates quotes for a different departure. There are 8 unique departure dates included in each graph. The dotted series indicates the aggregate value from all 8 departures.

2 Data Sources

The primary data for our analysis was collected using daily price quotes from a major travel search web site over the period February 22, 2011 to June 23, 2011. A web spider was written to query for each unique route and departure date pair in our study, so the results should be representative of what an individual user could observe in the market. Each query returned approximately 1,200 unique round-trip itineraries from all airlines; most queries returned results from 10 or more airlines. All itineraries were stored in a database, and feature values (discussed in Section 4) were computed as aggregates from the set of returned itineraries on each day. For consistency, these web queries were run sequentially and at the same time for every day in the study.[1]

Bing Travel, a popular travel search web site, has a "Fare Predictor" tool that provides a daily buy/wait policy recommendation for many route and departure dates. Bing Travel recommendation data was obtained from the Bing Travel search site for request date range March 15, 2011 to June 23, 2011. The site provides additional output of their model in the form of a distribution over future changes in the minimum price for each route and departure date over the next 7 days. We have collected these data in addition to our query data to facilitate comparisons between our model's buy/wait policy and the policy computed by the Bing Travel site. It is our understanding that the Bing Travel "Fare Predictor" is an extension of work by Etzioni et. al. in [5]. Examples of queries and statistics are shown in Table 1.

	Example 1	Example 2
Quote Date:	13 May 2011	13 May 2011
Origin City:	MSP	NYC
Destination City:	NYC	LAX
Departure Date:	20 May 2011	20 May 2011
Return Date:	25 May 2011	25 May 2011
No. of itineraries returned:	1135	1304
No. of airlines quoting:	9	13
No. of airlines exceeding 40% threshold:	8	10

Table 1: Airline price quote specification for the set of itineraries available for all airlines for specific 5-day round trips. The exact dates and cities shown are for illustration purposes only. The itinerary counts returned for these queries are also shown.

2.1 Pricing Patterns in Historical Data

Historical price quotes visible from a buyer's perspective can be used to develop predictive models of the sellers' pricing. There are strong cyclic patterns in the time series of prices. For example, Figure 1 shows the average lowest price quoted by all airlines for a specific origin-destination pair for 2 months of itineraries departing on (a) Thursdays and (b) Mondays. The Thursday to Tuesday time series shows a regular decrease in prices for Tuesday, Wednesday, and Thursday purchases,[2] while the (b) series shows significant increases for Thursday, Friday, and Saturday purchases. As expected, both series exhibit price increases in the last few days before departure (days to departure ≤ 7) but the (b) series exhibits this increase earlier in the time series. We posit that the majority of business flights would be Monday to Friday itineraries, and thus the demand for (b) series flights would be more insensitive to price than leisure flights. On business flights, airlines are able to increase prices sooner without causing a significant reduction in demand.

The pricing behaviors exhibited for other origin-destination pairs also differ from the example in Figure 1. A high traffic origin-destination pair such as the New York City to Los Angeles route exhibits much weaker

[1] All data sets used in our experiments are available, upon request, from the authors.

[2] *days-to-departure* **modulo** $7 \in \{0, 1, 2\}$

cyclic patterns. This example is shown in Figure 2. We conjecture that strategic pricing is likely to have a much greater observed effect for routes that have relatively few (2-3) competing airlines.

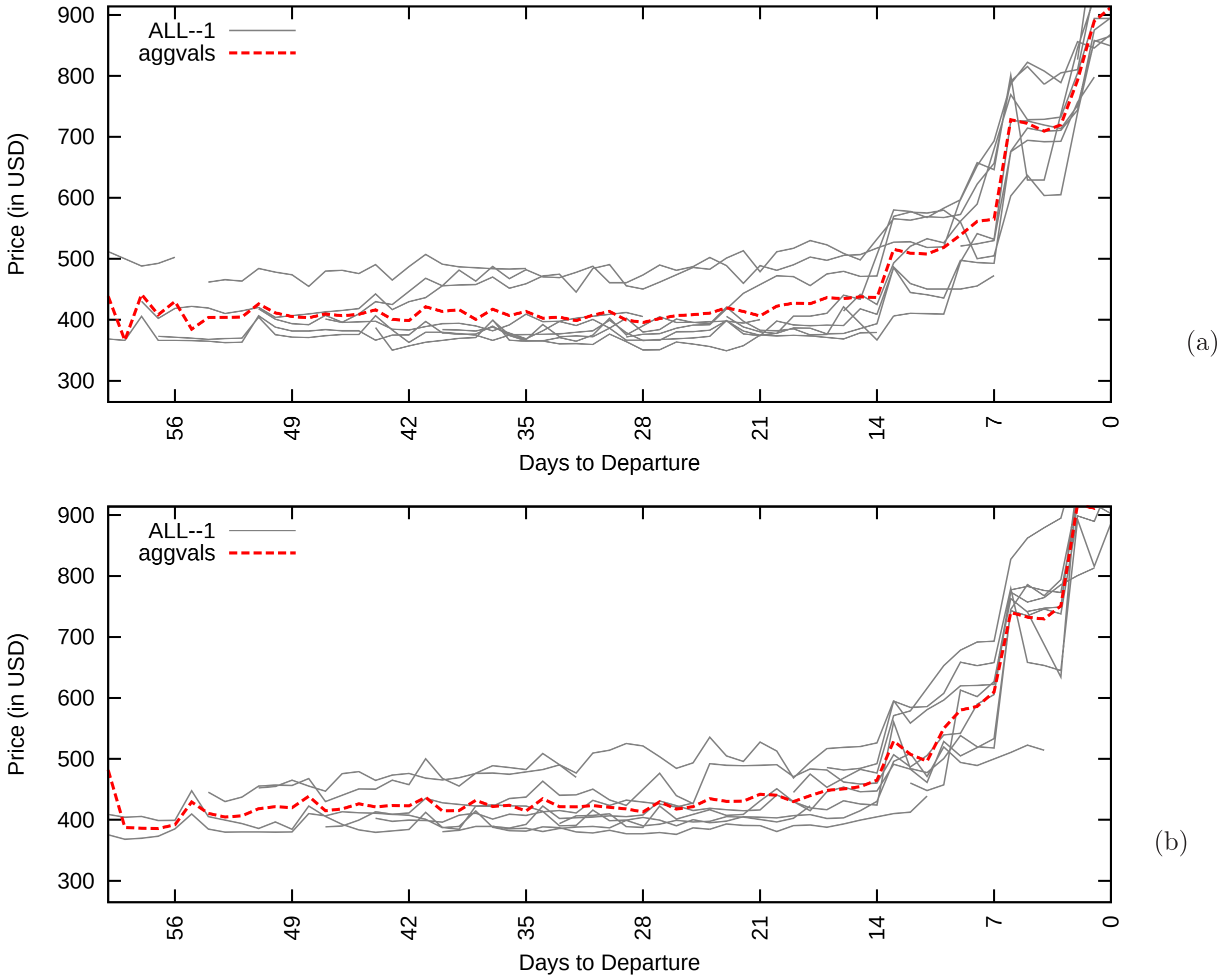

Figure 2: Mean lowest price offered by all airlines for NYC to LAX 5-day round trip flights having (a) Thursday departure and Tuesday return, or (b) Monday departure and Friday return itineraries.

3 Background and Related Work

Airlines determine the prices to offer for each flight through a process called yield management which is designed to maximize revenue given constraints such as capacity and estimates of future demand. For an overview of this process and the techniques used, see [2, 12]. Airlines divide seats in each flight into fare classes and charge different prices for each class in order to maximize overall revenue on their entire flight network. Airlines must balance the rate at which seats fill on their flights and this is generally done by changing the price. There are generally mismatches between airplane size and passenger demand are equalized through pricing, which has the effect of adjusting demand. Choosing optimal pricing on an entire airline network becomes increasingly complex because there are instances (in hub-and-spoke networks) when sacrificing revenue on a particular flight can increase overall revenue of the entire network.

The current state of yield management and competition in the airline industry is a direct result of historical decisions made about regulation in the industry [12]. The authors of [12] provide a case study describing the evolution of yield management at American Airlines. The techniques evolved beginning with a simple overbooking of flights. during the period of airline regulation (ending in 1979). Due to regulatory changes, airlines became free to adjust the airfare for each seat without restriction. This allowed airlines to divide the seats for each flight into different "fare classes" and charge different prices for effectively the same service. The development of fare classes was critical in maximizing passenger throughput in hub-and-spoke air networks because a passenger taking a single non-stop flight will accrue a different amount of revenue than a passenger taking a longer multi-stop flight. To maximize revenue, an airline needs to be able to offer competitive fares to both types of passengers and yield management is a way to amortize these differences within the company.

In traditional yield management, the lowest air ticket prices quoted to customers are based on the available seats in each fare class for a particular flight (or origin-destination pair in the case of multi-stop itineraries). An airline can adjust the rate-of-fill for a particular flight by moving seats between fare classes (i.e. by moving high cost seats into lower cost fare classes). These decisions are traditionally made by humans who take into account previous demand, current sales, and competitive market conditions.

Yield management can be applied to other industries with properties such as the need to handle advance reservations, a range of customer values for the same product, the ability for customers to cancel, a non-negligible probability of no-shows, or stock perishable inventory [4]. Industries with these properties include hotel booking, railroad transport (linear networks, many origin-destination pairs along a shared linear route), car rental, electric utility, and broadcasting industries. Dynamic pricing can also be beneficial in industries that can store inventory but these techniques have traditionally not been applied because of the high cost of changing prices. The key features enabling dynamic pricing are availability of demand data, ability to inexpensively change prices, and availability of decision support tools.

Additional market studies have addressed how the airline market has changed with the introduction of low cost airlines (LCAs). An overview of the competitive considerations in pricing strategies developed by LCAs in the European air travel market is in [11]. A general econometric model (using ordinary least squares regression) is developed to assess the most significant factors determining ticket prices from LCAs. The authors find that tickets purchased between 30 and 8 days prior to departure are more expensive than tickets bought in other periods. Tickets bought in the few days prior to departure can be significantly cheaper but are not always available due to demand. It should be noted that the LCAs do not compete against conventional airlines on price alone. They also use horizontal product differentiation to minimize the necessity to compete on price both with other LCAs and with conventional airlines. Specifically, LCAs attempt to locate themselves at secondary airports (not significantly served by conventional airlines) and fly on schedules that are maximally distant from existing players. This suggests that preferences about schedule convenience and location also play a significant role in customers' purchasing decisions and ought to be considered in any predictive model in this domain.

In a later investigation [1] on measurements of market power of LCAs in the European airline market, airlines that have a significant share of the traffic at an individual airport tend to have higher prices than other carriers at the same airport. Also, an airline having a large portion of traffic between two pairs of airports (one direct route) tends to have greater market power than an airline having a large portion of traffic between two airports without a direct route. There is greater substitutability on routes with one or more stops, so market power is lower.

Some work has been done on determining optimal purchase timing for airline tickets. Our work has been inspired by [5], where several purchasing agents attempt to predict the optimal purchase time of an airline ticket for a particular flight. The models are able to determine the optimal purchase time within the last 21 days prior to departure for specific flights in their collected data set. For benchmarking the results, the authors compute the purchasing policy (a sequence of wait/buy signals) for many unique simulated passengers with a specific target airline, target flight, and date of departure to satisfy. The optimal policy (the sequence of buy/wait signals that leads to the lowest possible ticket price) is used as a benchmark for each simulated passenger and the cost of each alternative purchasing agent is computed. The aggregate result shows that, given these purchasing criteria, it is possible to save a significant amount when purchasing. This paper is different from the previous work in that we model the aggregate cost of all flights meeting some preference criteria.

There are several efforts in the game theory community to model aspects of the airline ticket domain, usually for the purpose of understanding competitive market dynamics of the oligopoly of sellers. In [13], a dynamic programming model is presented for determining optimal fare class allocation (of 4 fare classes) on a single flight. This model incorporates fare class-dependent and time-dependent cancellation, overbooking, and no-show probabilities. Valuable insights provided by this study are that booking limits need not change monotonically over time (may increase or decrease), it may be optimal to accept a lower fare class while simultaneously rejecting a higher fare class (due to differences in cancellation characteristics), and cancellations cause the optimal policy to depend on both total capacity and remaining capacity. The critical disadvantage of this approach is that extending it to even a small size airline network would result in a significant increase in complexity due to the booking interactions between multi-leg flights.

A one-shot game theory-based simulation of pricing competition in the airline ticket price domain is presented in [7]. When two airlines with significant capacity compete with each other and their products are not sufficiently differentiated, the equilibrium price falls to a minimum price threshold, referred to as the “spiral down” price. This result may shed some light on the long term decisions airlines make about airplane size and flight frequency. The authors also note that the airline pricing domain is more similar to a repeated game than a one-shot game. Other equilibria can be enforced in repeated games that are significantly above the spiral down price found in the non-repeated game. This work also shows that a completely automatic pricing mechanism can be potentially ruinous for an airline. There must be supervisory mechanisms that take into account other aspects into pricing beyond price competition.

A game theory model of dynamic pricing that incorporates an oligopoly facing strategic customers, buyers who will delay purchase until a future time period if there is a high likelihood of obtaining a lower price later, is presented in [9]. The work assumes perfect foresight, all parties (sellers and customers) can estimate perfectly future outcome probabilities and utilities. If even a portion of the population of customers is strategic, revenue is reduced for the sellers and any strategic defenses in such a transparent market cannot fully ameliorate this effect. These findings were found to persist among different oligopolies of sellers including monopoly, duopoly and three-seller oligopolies. In particular, the disadvantages for sellers in the presence of strategic customers increased as competition (number of selling firms) increased. The critical conclusion of this work is that the most effective method to inhibit the impact of strategic consumers is to reduce the amount of information available to consumers. This may explain in part why, in spite of the technical feasibility, few significant predictive tools have been made available to individual purchasers of airline tickets.

4 Our Model

Feature Extraction. The number of itineraries (¿1000) in each daily query made some aggregation necessary. The features extracted for prediction are aggregated variables computed from the (large) list of quotes observed on individual query days. For each query day, there are possibly many airlines quoting flights for a specific origin-destination and date combination. However, not all airlines will quote every day. This is possibly due to strategic decisions of the airline or due to lack of available capacity. We limit the number of airlines used for distinct features by focusing on airlines that quote for a specific route more than 40% of the query days. Also, each airline may present itineraries that contain non-stop segments or segments with one or more stop. We separate the quotes by their number of stops into three bins: non-stop round trips, round trips with a maximum of one stop in each direction, and round trips with 2 or more stops in either direction. For each bin, three features are computed: the minimum price, mean price, and the size of the bin (the number of quotes). Additionally, these three features are computed for the union of all three bins. So for each airline, 12 features are computed on each quote day. For airlines that do not exceed the 40% of quote days criteria, their itineraries are combined into a separate “OTHER” category placeholder for which the same 12 features are generated. Finally, these same 12 aggregates are generated for all itineraries returned and are placed in the “ALL” airlines category. An additional computed feature, referred to as *days to departure*, is the number of days between the quote date and the departure date. On a quote date where a

Class D0 (no. of vars.: 8)	Class A1 (no. of vars.: 3)	Class A2 (no. of vars.: 9)	Class A3 (no. of vars.: 18)	Class A4 (no. of vars.: 54)
Days to departure	ALLminpA[a]	ALLminp0[b]	aDLminpA[c]	aDLminp0
Quote Day-of-week is Mon.[d]	ALLmeanpA	ALLmeanp0	aDLmeanpA	aDLmeanp0
Quote Day-of-week is Tues.	ALLcountA	ALLcount0	aDLcountA	aDLcount0
Quote Day-of-week is Wed.		ALLminp1	...	aDLminp1
Quote Day-of-week is Thurs.		ALLmeanp1	OTHERminpA	aDLmeanp1
Quote Day-of-week is Fri.		ALLcount1	OTHERmeanpA	aDLcount1
Quote Day-of-week is Sat.		ALLminp2	OTHERcountA	aDLminp2
Quote Day-of-week is Sun.		ALLmeanp2		aDLmeanp2
		ALLcount2		aDLcount2
				...
				OTHERcount2

[a] minimum price quoted by any airline, for any number of stops
[b] minimum price quoted by any airline, for non-stop flights only
[c] minimum price quoted by a specific airline (DL = Delta Airlines), for any number of stops
[d] 1 if quote is from a Monday, otherwise 0

Table 2: Raw features (sorted by feature class) available for each quote day for a specific departure day and route. The precise number of features in some classes (A2, A4) will vary based on the number of airlines quoting the route. The variable counts given are specific to the MSP-NYC route (92 total raw features).

specific airline does not quote any flights matching the criteria, the value from the aggregate "ALL" airlines category is used. A listing of the variables computed on each quote day is shown in Table 2. Using the computed feature vectors and the corresponding values for some target variable, a machine learning problem can be formulated for performing prediction.

Policy Computation and Evaluation. As a first step, the aggregated features computed above can be used by a regression model for prediction of the expected lowest price for a specific round trip between the quote day and departure. Such a model can determine if the currently quoted price is, relatively speaking, a bargain. An obvious approach is to choose the regression model with the best accuracy of all candidates, but it may not be the model that generates the lowest average cost policy. A better way to measure performance of this kind of prediction model is to measure performance (cost) that results from following the computed policy recommendation. To that end, we couple each prediction model output $\hat{e}_t$, the model estimate future price at time t, with a decision threshold d, either an absolute price difference $d_{\$}$ (in \$) or a relative price difference $d_{\%}$ (a percentage), and compare against the current day's price p_t to determine the policy recommendation (buy or defer purchase) for each day and flight.

Table 3 shows the relationship between the price model, the associated decision threshold, and the policy recommendation r_t. The prediction model is trained and accuracy is measured to achieve the highest possible accuracy as measured by root mean square error (RMSE). However, it is also important to discount the utility of future anticipated low prices (using a threshold) in order to determine the optimal level of risk. Our method determines the optimal threshold value d for each model that results in the lowest average ticket purchase price.

The following is an example of the decision threshold computation. A model trained to estimate the minimum price between the current day and departure for the MSP-NYC route[3] that is coupled with a decision threshold -5% would compute the following logic: if the current day's price is 5% lower than the prediction $\hat{e}_t$, then purchase today, otherwise defer purchase. The optimal decision threshold for each trained model can be determined by searching the range of possible values; the result of this search is shown in Section 5.

We tested two types of decision thresholds in our experiments: (1) a percent based threshold and (2) a

[3] The results of this model computation are called "ProposedM: Dep" in Table 7.

Type	Formula
Percentage	$r_t = \begin{cases} \text{WAIT} & : \hat{e}_t < p_t(100\% + d_{\%}) \\ \text{BUY} & : \text{otherwise} \end{cases}$
Absolute	$r_t = \begin{cases} \text{WAIT} & : \hat{e}_t < p_t + d_{\$} \\ \text{BUY} & : \text{otherwise} \end{cases}$

Table 3: Policy computation by decision threshold type for day t using model estimate $\hat{e}_t$, decision threshold d, and current ticket price p_t.

relative price based threshold. We expect both threshold types to produce a good result but there may be slight differences in the performance of the two methods. The percentage based threshold should produce more consistent results if the price of a particular flight varies dramatically over the period of interest. The range of $d_{\%}$ values searched was a percentage range {-30, 30} in integer increments, and the range of $d_{\$}$ searched was {-100, 100} in increments of 5 dollars. The result of this search for developing the purchase policy are shown for each target in the results section.

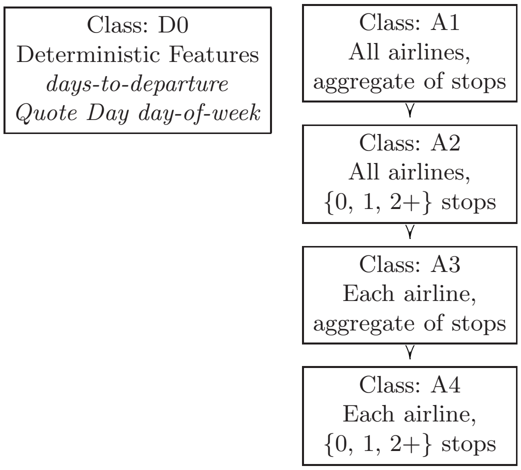

Figure 3: Lag scheme class hierarchy for product price prediction. Arrow denotes a *not greater than* relationship (i.e. class A1 should have an equal or higher maximum time offset than class A2).

Lagged feature computation. Using only the most recent values (92 features for the MSP-NYC route) as the entire feature set may provide reasonable prediction results in some domains, but such a model cannot predict trends or temporal relationships present in the data. The need to represent temporally-offset relationships motivates the idea of adding time-delayed observations to the feature set as well. We refer to this as the addition of *lagged features*. For instance, if the cost of a route on day $t - 7$ is representative of the price of a available on day $t + 1$, the 7 day delayed observation should have a high weight in the model.

Our technique uses the assumption that more recent observations are likely to have high informational value for price prediction, but time-delayed features may hold informational value as well (i.e. the environment is not completely stochastic). The time between a change in the market and its effect on the target variable may be longer than one day and lagged variables can leverage those delayed relationships. Even with the constraint that more recent observations are added before less recent observations, searching for the optimal subset from 92 available features is still an intractably large search space.

To reduce the number of possible configurations, we introduce the notion of a hierarchical segmentation of the feature set. In this domain, each of the 92 features is placed into one of several classes based on its specificity using a minimal amount of domain knowledge as shown in Figure 3.

By searching through all combinations (a small number) of feature classes used it is possible to automatically tune the final feature vector for each target regression. Another novelty of our method is to allow the addition of time-delayed instances of each class to facilitate the prediction of trends. An additional constraint is added: more specific classes will not have more time delayed instances than more general classes.

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●		
A2	●			
A3				
A4				

(a) All Airlines

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●			
A3				
A4				

(b) All Airlines, non-stop

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●	●	
A3				
A4				

(c) Airtran Airlines

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●		
A3	●	●		
A4	●	●		

(d) American Airlines

Table 4: Optimal lag schemes for various preferences on the minimum overall ticket price on route MSP-NYC 5-day round-trip with Thursday departure.

The simplest lag configuration, shown in Table 4, contains the most recent day's value from all feature classes. We posit that time-delayed observations from the variable of interest (such as the all airline minimum price in *Class A1*) are likely to be predictive as well. Time-delayed observations from other more-specific feature classes *may also be* but are less likely to be predictive. It is by this principle that the hierarchy and strict ordering of lagged data additions is based. By constraining the classes so that the less informationally dense classes have lower time delays and contribute fewer additional features, we prevent the inclusion of extraneous, irrelevant features.[4] Additional examples of optimal lag schemes for different targets is shown in Table 11 of the appendix.

Next, we will show how the time lagged data is constructed to form the augmented features set. An expansion of the feature set is referred to as a *lag scheme expansion*. Of course, the optimal lag scheme may be different for each variable modeled; a search of the possible configurations is performed to find the best performing configuration for a target.

The number of possible lag schemes as formulated with the hierarchy in Figure 3 for a maximum time delay of 7 days is 108. Without the constraints between classes, there are 1250 configurations of the 5 feature classes if constrained to possible time delays of $\{\varnothing, 0, 1, 2, 7\}$, but many of these configurations will be uninteresting variants. Finally, without the hierarchical segmentation and constraints between classes, there are $\approx 10^{62}$ configurations of the 92 original features.[5] Using both the feature classification and the constraint hierarchy allows for a greater variety of "interesting" lag schemes to be tested for the same amount of search.

While a domain expert could conceive of a generally high-performance feature set, the automated lag scheme search should contain a configurations similar to what a domain expert can build (without the cost of becoming a domain expert). Also, the results of the optimal lag scheme search can elicit some surprising

[4] Days-to-departure and the day of the week are deterministic features, their value from one time-offset can be computed deterministically from another time-offset's value, so there is no predictive value in including these more than once.

[5] 84 price features, and 8 deterministic features (days-to-departure and quote day of the week) $= 2 * (2^7) * (5^{84})$

relationships found in the data.

Table 4 shows the optimal lag schemes for several targets. It is interesting to note that more specific preferences (b,c,d) benefit from a larger feature set (both in temporal depth and feature class breadth).

Model Construction using PLS Regression. The novelty of our approach does not rely on any modifications to the PLS algorithm, so our treatment of PLS will be brief. Mathematically, PLS regression deterministically computes a linear function that maps a vector of the input features x_i into the output variable y_i (the label) using a vector of weights $\bar{w}$. Several implementations of PLS exist [3, 14, 10]; each with its own performance characteristics. This work uses the orthogonalized PLS, Non-Integer Partial Least Squares (NIPALS), implementation in [14]. PLS was chosen over similar multivariate techniques including multiple linear regression, ridge regression [6], and principal component regression [8] because it produces better performance than the others and has an ability to adjust model complexity.

This algorithm has multiple advantages. First, PLS regression is able to handle very high-dimensionality inputs because it implicitly performs dimensionality reduction from the number of inputs to the number of PLS factors. Second, the model complexity can be adjusted by changing the number of PLS factors to use in computing the regression result. This value is adjusted in our experiments to determine the optimal model complexity in each prediction class. Third, the algorithm is generally robust to highly collinear or irrelevant features. Fourth, the structure of a trained model can be examined for knowledge about the domain. For these reasons, this algorithm was chosen.

The PLS regression algorithm in [14] allows users to adjust the model complexity by selecting the number of PLS factors to generate when training. (These factors are analogous to the principal component vectors used in principal component regression.) The number of PLS factors determines the dimensionality of the intermediate variable space that the data is mapped to. The computational complexity does not significantly increase for a larger number of factors but the choice does have an effect on prediction performance: too many factors can cause over-fitting, and too few factors can cause the model to be unable to represent relationships in the data. We systematically varied the number of factors in the optimal lag scheme search. The best performing number of components is shown for each prediction category in Section 5.

	Purchase Method (all values in $USD)			
	Immediate Purch.	Bing Travel (% savings over “Immed. Purch.”)	Our Method	Optimal
Any Airline	320	318 (0.70%)	296 (7.5%)	281 (12%)
Any Airline, non-stop	462	461 (0.1%)	416 (9.9%)	399 (13%)
Airtran Airlines	344	343 (0.2%)	333 (3.3%)	303 (12%)
American Airlines	460	462 (−0.2%)	416 (9.6%)	403 (12%)
Continental Airlines	469	468 (0.1%)	409 (13%)	403 (14%)
Delta Airlines	458	460 (−0.4%)	420 (8.3%)	403 (12%)
Delta Airlines, non-stop	544	546 (−0.4%)	509 (6.4%)	489 (10%)
Delta Airlines , 1 stop	499	501 (−0.4%)	474 (4.9%)	462 (7.4%)

Table 5: Compares purchase algorithms by the average minimum cost (in $) for 5-day round-trip Thursday departure tickets by specific airlines bought less than 60 days prior to departure from MSP to NYC.

5 Experimental Results

Our experiments were designed to estimate the differences in costs of using our prediction models to develop a purchase policy. The prices for airline tickets more than 60 days prior to departure tend to vary infrequently and not in significantly large price movements. Also, the literature suggests that demand for flights more than two months in advance is relatively low. While precise details of the evolution of demand for a particular flight are proprietary secrets in the airline business, some information has been published on the expected demand profile over time that are used by airlines in their own pricing models: first, airlines assume a relatively fixed rate of purchases until a flight is full, and second, most tickets for a flight are sold within 60 days of departure [2, 12]. Using these facts, we conjecture that a good performance measure for a purchasing strategy would be to compute the cost of following the purchase recommendations for an itinerary once for each day in the range of α days to 1 day prior to departure. In these experiments, α is set to 60. This measure involves hypothetically purchasing an itinerary precisely α times for each purchase algorithm under test (but some purchases may be deferred for a few days based on the recommendation of the algorithm). Each of the α purchases is called a *purchase episode*.

Performance Measure. The most basic purchase algorithm, called *immediate purchase*, is to purchase a ticket once for each day in the α day range. All purchase episodes of this type would terminate with a purchase event on the first day of the episode and cost would be equal to the sum of prices observed in the α day period. The lowest achievable cost is called the *optimal cost* and is based on purchasing for each of the α episodes at the lowest price observed between the beginning of the episode and the departure date. Table 6 provides examples of several purchase episodes for one target, the policies and the associated costs. One would expect that the best purchase policy algorithm would achieve an overall cost somewhere between the *optimal cost* and *immediate purchase* methods but could in the worst case achieve costs higher than the latter. In Table 5, we supply the results of estimated costs for several purchasing policies based on purchasing 4 different 5-day Thursday to Tuesday itineraries from MSP to NYC (a total of 240 purchases per method). The training set for this period consisted of the preceding 4 weeks of departures. We also show how costs vary based on preferences such as a customer requiring specific airline or a specific number of intermediate stops (i.e. non-stop, 1 stop). We also compare our best policy result against the cost of following the buy/wait recommendation from Bing Travel's "Fare Predictor." A detailed comparison is shown in Table 7. The comparison methodology used here involving simulated purchases that follow a model-computed policy is similar to that used in [5].

Days to depature		39	38	37	36	...	2	1	0
naïve	action	•	•	•	•	...	•	•	•
	cost	262	262	258	257	...	385	453	453
bing travel	action	◯	•	•	•	...	•	•	•
	cost	262	262	258	257	...	385	453	453
optimal	action	◯	◯	◯	◯	...	•	◯	•
	cost	171	171	171	171	...	385	453	453
time series model	action	◯	◯	◯	◯	...	•	◯	•
	cost	229	229	229	229	...	385	453	453

Table 6: Example of policies computed by various models. The ◯ symbol indicates a "WAIT" signal for that day, and the • symbol indicates a "BUY" signal for that day. The models are compared elsewhere with each other by considering the mean value of the cost vector shown for each model.

Comparing the immediate purchase cost with the optimal purchase cost shows that there is, on average, a possible 10% savings to be achieved over *immediate purchase*, the most naïve approach. We denote this percentage as the *savings margin*. Our method is able to achieve a savings margin in the range of 5%, about half the distance between the naïve and the optimal.

Bing Travel Performance Comparison. It may be unsurprising that the Bing Travel recommendations do not save significantly over the immediate purchase recommendations for the airline specific target variables. The website only advertises that it provides a prediction about whether or not the lowest cost ticket from any airline will be lower over the next 7 days. By this criteria, it is only possible to forecast about the weekly cycles seen in the price time series. Also, the price changes of an individual airline do not necessarily follow the pattern of the lowest cost price, so using these recommendations in this way is not entirely valid. For this reason, those values are shown in gray in Table 5. It is surprising however that Bing Travel is not able to achieve a greater savings margin on the Any Airline target. We posit that this is due to a generally risk averse approach taken by their algorithm: it is more much more likely than our method to advise immediate purchase than it is to advise waiting. This assertion can be validated by looking at the distribution of buy and wait signals computed for each day by the various policy generators.

City Pair	Trip	Model	Buy Signal (%)	Wait Signal (%)	Mean Wait (days)	Wait Std. Dev. (days)	Mean Cost ($)	Cost Std. Dev. ($)	Decision Threshold	Lag Scheme
MSP-NYC	M-F	naive	100	0	0.00	0.00	296	60.9	–	–
		optimal	11	88	10.3	9.57	249	79.7	–	–
		Bing Travel[a]	83	17	0.349	0.954	293	61.5	–	–
		ProposedM:7[b]	14	86	8.60	8.80	262	75.8	+2%	78
		ProposedM:D[c]	23	77	5.88	6.04	265	77.0	−3%	98
	Th-Tu	naive	100	0	0.00	0.00	278	43.0	–	–
		optimal	11	89	11.0	9.71	221	54.4	–	–
		Bing Travel	76	23	0.540	1.23	271	43.2	–	–
		ProposedM:7	23	77	7.39	8.27	242	52.1	+$10	3
		ProposedM:D	17	83	8.24	8.69	243	57.1	−7%	55
NYC-LAX	M-F	naive	100	0	0.00	0.00	353	74.0	–	–
		optimal	16	83	7.74	8.09	306	94.4	–	–
		Bing Travel	73	27	0.785	1.86	443	88.0	–	–
		ProposedM:7	12	87	8.63	7.92	314	113	−5%	0
		ProposedM:D	14	85	8.34	8.01	316	104	−$25	54
	Th-Tu	naive	100	0	0.00	0.00	333	61.1	–	–
		optimal	17	83	7.09	7.04	302	75.7	–	–
		Bing Travel	70	30	0.817	1.69	378	57.9	–	–
		ProposedM:7	27	73	4.04	4.76	309	95.8	−2%	98
		ProposedM:D	30	70	4.37	5.44	304	88.2	−2%	98

Table 7: Buy/Wait policy recommendation distribution comparison by model for city pairs MSP-NYC and NYC-LAX, and 5-day trip periods Monday-Friday and Thursday-Tuesday.

[a] Bing Travel's buy or wait policy reccommendation.

[b] The propsed model used to estimate the minimum price for 7 days into the future.

[c] The proposed model used to estimate the minimum price for all days until departure.

Table 7 compares the raw number of buy and wait signals output by each model in our study. It is perhaps unsurprising that the optimal policy has a high proportion of wait signals: in the MSP-NYC M-F route, the optimal policy has only 11% proportion of buy signals. It is noteworthy that the best models constructed with our method also emits a similar proportion of wait signals: in the MSP-NYC M-F route, the model with the lowest average cost ($262) only emits a buy signal on 14% of the days in the test set. The Bing Travel model has a much higher proportion of buy signals: in the same route, the Bing Travel model emits a buy signal 83% of the days. The results are similar for all routes and dates in our survey: Bing emits buy signals for at least 70% of the days. While the precise reasons for the Bing Travel model bias towards buy signals is unknown, we posit that the model may be more averse to future possible price increases than our tuned minimum cost approach.

	Significant Competitors[6]	Mean Passengers per Day	Std. dev. of Lowest Cost Ticket by Departure[7]	
			Thu.	Mon.
MSP → NYC	3	1012	0.058	0.065
NYC → LAX	6	4031	0.035	0.043

Table 8: Comparison of the two airline routes under investigation.

Significant Competitors. Comparing the effectiveness of our policy construction approach on two different airline routes has led to some insights about differences in the market structure of the two routes. By comparing the results of MSP-NYC models against the NYC-LAX models in Table 7, reveals that there is a smaller savings margin in the NYC-LAX data. The price quotes reveal that there is significantly more competition and passenger volume in the NYC-LAX route. A comparison[8] of the two routes across several measures is provided in Table 8. The lower variance of daily minimum prices shown in the NYC-LAX route is likely due to the large number of competitive carriers along the route. In contrast, the MSP-NYC route has fewer "significant competitors", so individual airlines can assert greater pricing power (and cause prices to fluctuate more strongly).

6 Conclusions and Future Work

This investigation shows that, given sufficient publicly-observable information, it is possible to predict airline ticket prices sufficiently to reduce costs for customers. We believe that there is a significant market for this these kinds of models in the hands of consumers. In particular, reliable price models can assist buyers in determining the range of expected prices for a particular itinerary. The current market environment does not provide customers with any reliable estimates of the future costs of any particular departure meeting their requirements. At best, frequent travelers can develop models of prices from their own past history.

In spite of being the most obvious purchase policy, buying at the earliest opportunity is not the best policy for most customers. First, the long lead-time price may not be the lowest price available for a particular flight before departure. Also, there is an opportunity cost associated with early commitment: a customer risks being locked into a specific schedule that may need to be changed (for a fee).

The novelty of this work rests in a regression model formulation for domains having significant intra-variable and inter-variable temporal relationships. The hierarchicalization of the feature set is possible with some domain knowledge, but expert level understanding is not required. The resulting lag scheme model can be examined for domain understanding.

Because there is sufficient structured price volatility on many airline routes, there are significant opportunities for saving money when purchasing by using the guidance of a predictive model. In addition to the results of this work, we believe there are additional cost reductions that can be found to obtain results closer to the optimal policy.

References

[1] E. Bachis and C. A. Piga. Low-cost airlines and online price dispersion. *International Journal of Industrial Organization*, In Press, Corrected Proof:–, 2011.

[2] P. P. Belobaba. Airline yield management. an overview of seat inventory control. *Transportation Science*, 21(2):63, 1987.

[3] S. de Jong. Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251 – 263, 1993.

[4] W. Elmaghraby and P. Keskinocak. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science*, 49(10):pp. 1287–1309, 2003.

[5] O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates. To buy or not to buy: mining airfare data to minimize ticket purchase price. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 119–128, 2003.

[8] The raw data source for computing these values was the US Department of Transportation Bureau of Transportation Services Origin-Destination Survey (see `www.bts.gov`).

[6] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):pp. 80–86, 2000.

[7] K. Isler and H. Imhof. A game theoretic model for airline revenue management and competitive pricing. *Journal of Revenue & Pricing Management*, 7(4):384–396, 2008.

[8] I. T. Jolliffe. A note on the use of principal components in regression. *J. of Royal Statistical Society. (Applied Statistics)*, 31(3):pp. 300–303, 1982.

[9] Y. Levin, J. McGill, and M. Nediak. Dynamic pricing in the presence of strategic consumers and oligopolistic competition. *Management Science*, 55(1):32–46, 2009.

[10] H. Martens and T. Næs. *Multivariate Calibration.* John Wiley & Sons, July 1992.

[11] C. Piga and N. Filippi. Booking and flying with low-cost airlines. *International Journal of Tourism Research*, 4(3):237–249, 2002.

[12] B. C. Smith, J. F. Leimkuhler, and R. M. Darrow. Yield management at American Airlines. *Interfaces*, 22(1):8–31, 1992.

[13] J. Subramanian, J. Stidham, Shaler, and C. J. Lautenbacher. Airline yield management with overbooking, cancellations, and no-shows. *Transportation Science*, 33(2):147–167, 1999.

[14] S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the PLS method. In *Matrix Pencils*, volume 973 of *LNM*, chapter 18, pages 286–293. Springer, 1983.

Appendix

The appendix contains tables with additional details from the experiments presented in this paper. In many cases, the tables in the paper are condensed versions of those shown here.

Table 9 shows a sample of the feature vector for an individual city pair. The exact number of features for an origin-departure pair will vary based on the number of airlines quoting more than 80% of the quote days. For example, if 5 airlines exceed the threshold, the number of features would be 92.

Table 10 compares various purchase algorithms by the average minimum cost (in $) for a 5 day round trip with Thursday departure by specific airlines bought less than 60 days prior to departure from MSP to NYC between March 12, 2011 and May 12, 2011.

Finally, Table 11 shows the optimal lag schemes by selected airlines and classes.

Feature	Description
Days to departure	Number of days between quote and departure dates
Quote Day-of-week is Monday	1 if quote is from a Monday, otherwise 0
Quote Day-of-week is Tuesday	1 if quote is from a Tuesday, otherwise 0
Quote Day-of-week is Wednesday	1 if quote is from a Wednesday, otherwise 0
Quote Day-of-week is Thursday	1 if quote is from a Thursday, otherwise 0
Quote Day-of-week is Friday	1 if quote is from a Friday, otherwise 0
Quote Day-of-week is Saturday	1 if quote is from a Saturday, otherwise 0
Quote Day-of-week is Sunday	1 if quote is from a Sunday, otherwise 0
ALLminpA	All airlines, minimum price
ALLmeanpA	All airlines, mean price
ALLcountA	All airlines, no. of itineraries
ALLminp0	All airlines, minimum price for non-stop itineraries
ALLmeanp0	All airlines, mean price for non-stop itineraries
ALLcount0	All airlines, no. of itineraries for non-stop
ALLminp1	All airlines, minimum price for 1 stop itineraries
ALLmeanp1	All airlines, mean price for 1 stop itineraries
ALLcount1	All airlines, no. of itineraries for 1 stop
ALLminp2	All airlines, minimum price for 2+ stop itineraries
ALLmeanp2	All airlines, mean price for 2+ stop itineraries
ALLcount2	All airlines, no. of itineraries for 2+ stop
aDLminpA	Delta Airlines, minimum price
aDLmeanpA	Delta Airlines, mean price
aDLcountA	Delta Airlines, no. of itineraries
aDLminp0	Delta Airlines, minimum price for non-stop itineraries
aDLmeanp0	Delta Airlines, mean price for non-stop itineraries
aDLcount0	Delta Airlines, no. of itineraries for non-stop
aDLminp1	Delta Airlines, minimum price for 1 stop itineraries
aDLmeanp1	Delta Airlines, mean price for 1 stop itineraries
aDLcount1	Delta Airlines, no. of itineraries for 1 stop
aDLminp2	Delta Airlines, minimum price for 2+ stop itineraries
aDLmeanp2	Delta Airlines, mean price for 2+ stop itineraries
aDLcount2	Delta Airlines, no. of itineraries for 2+ stop
... additional airlines ...	
OTHERminpA	Other Airlines, minimum price
OTHERmeanpA	Other Airlines, mean price
OTHERcountA	Other Airlines, no. of itineraries
OTHERminp0	Other Airlines, minimum price for non-stop itineraries
OTHERmeanp0	Other Airlines, mean price for non-stop itineraries
OTHERcount0	Other Airlines, no. of itineraries for non-stop
OTHERminp1	Other Airlines, minimum price for 1 stop itineraries
OTHERmeanp1	Other Airlines, mean price for 1 stop itineraries
OTHERcount1	Other Airlines, no. of itineraries for 1 stop
OTHERminp2	Other Airlines, minimum price for 2+ stop itineraries
OTHERmeanp2	Other Airlines, mean price for 2+ stop itineraries
OTHERcount2	Other Airlines, no. of itineraries for 2+ stop

Table 9: Feature vector example

	Model												Immed. Purch. Cost (in $)
	7 day minimum price				until dep. minimum price				time series				
	lag scheme no.	decision threshold	complexity (# of features)	savings margin over immed. purch.	lag scheme no.	decision threshold	complexity (# of features)	savings margin over immed. purch.	lag scheme no.	decision threshold	complexity (# of features)	savings margin over immed. purch.	
Any Airline	56	+ 4%	6	7.5%	55	− 6%	6	8.9%	59	+ 4%	6	9.5%	320
Any, non-stop	89	− 25	6	7.5%	89	− 50	6	7.6%	67	+ 3%	6	8.3%	462
Airtran	62	+ 0%	6	5.9%	98	− 6%	6	6.0%	107	+ 8%	3	6.8%	344
American	97	− 5	6	18.4%	97	− 20	6	20.8%	97	+ 5%	3	17.5%	460
Continental	33	− 15	6	10.9%	78	− 8%	6	11.9%	63	+ 4%	6	12.4%	469
Delta	52	+ 4%	6	13.7%	53	− 10	6	14.4%	42	+ 5	6	11.1%	458
Delta, non-stop	92	+ 2%	6	12.4%	98	− 5	6	13.1%	67	+ 3%	6	11.5%	544
Delta, 1 stop	92	+ 5	6	12.3%	92	− 5	6	12.3%	67	+ 5	6	10.9%	499

Table 10: Comparison of various purchase algorithms by the average minimum cost (in $) for (5 day round trip Thursday departure) tickets by specific airlines bought less than 60 days prior to departure from MSP to NYC.

Model: Min. Price Next 7 Days

Model: Min. Price Until Departure

Model: Time Series

All Airlines

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●			
A2	●			
A3	●			
A4				

Lag Scheme 56

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●			
A2	●			
A3				
A4				

Lag Scheme 55

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●		
A2	●			
A3				
A4				

Lag Scheme 59

All Airlines non-stop

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●			
A3				
A4				

Lag Scheme 89

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●			
A3				
A4				

Lag Scheme 89

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●		
A2	●	●		
A3	●	●		
A4	●	●		

Lag Scheme 67

Airtran Airlines

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●		
A2	●	●		
A3				
A4				

Lag Scheme 62

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●	●	
A3				
A4				

Lag Scheme 98

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●	●	
A3	●	●	●	
A4	●	●	●	

Lag Scheme 107

American Airlines

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●		
A3	●	●		
A4	●	●		

Lag Scheme 97

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●		
A3	●	●		
A4	●	●		

Lag Scheme 97

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●		
A3	●	●		
A4	●	●		

Lag Scheme 97

Continental Airlines

Class	Lagged Offsets			
	0	1	2	7
D0				
A1	●	●	●	
A2	●	●	●	
A3	●	●	●	
A4	●	●	●	

Lag Scheme 33

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	
A2	●	●	●	
A3				
A4				

Lag Scheme 78

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●		
A2	●	●		
A3	●			
A4				

Lag Scheme 63

Delta Airlines

Class	Lagged Offsets			
	0	1	2	7
D0				
A1	●	●	●	●
A2	●	●	●	
A3	●	●	●	
A4	●	●		

Lag Scheme 52

Class	Lagged Offsets			
	0	1	2	7
D0				
A1	●	●	●	●
A2	●	●	●	
A3	●	●	●	
A4	●	●	●	

Lag Scheme 53

Class	Lagged Offsets			
	0	1	2	7
D0				
A1	●	●	●	●
A2	●	●		
A3	●	●		
A4	●			

Lag Scheme 42

Delta non-stop

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●		
A3				
A4				

Lag Scheme 92

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●	●	
A3				
A4				

Lag Scheme 98

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●		
A2	●	●		
A3	●	●		
A4	●	●		

Lag Scheme 67

Delta 1-stop

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●		
A3				
A4				

Lag Scheme 92

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●	●	●
A2	●	●		
A3				
A4				

Lag Scheme 92

Class	Lagged Offsets			
	0	1	2	7
D0	●			
A1	●	●		
A2	●	●		
A3	●	●		
A4	●	●		

Lag Scheme 67

Table 11: Optimal lag schemes by selected variables (airline and number of stops) and model targets. Class D0 contains *days-to-departure* and *quote day day-of-week* features. Class A1 contains the 3 aggregate features (minimum price, mean price, and number of quotes) computed from all quotes on each day. Class A2 contains the 3 aggregate features broken apart by number of stops: 0, 1, and 2+. Class A3 contains the 3 aggregate features computed for each airline. Class A4 contains the 3 aggregate features computed for each combination of airline and number of stops.